УДК 80

Амиева А. М., *Филимонов В. В., Сергеев А. П.*
УрФУ, г. Екатеринбург, Россия

ОСНОВНЫЕ МЕТОДИКИ ИССЛЕДОВАНИЯ СТРУКТУРЫ ТЕКСТА

Аннотация

Проблема обнаружения скрытых структур текста связана с перспективной методикой установления авторства. В работе обосновывается выбор методики для обнаружения и изучения таких структур в русскоязычных текстах. Рассмотрены основные методики, оперирующие в качестве единиц анализа смыслосодержащими элементами, лексическими или графическими формами и частотными характеристиками текста. Авторы предполагают существование скрытых структур, исходя из ранее полученных результатов. Предлагается использование частотного анализа как наиболее адекватного задаче обнаружения скрытых структур текста.

Ключевые слова: текст, структура, методика, семиотика, частотный анализ.

Amieva A. M., *Filimonov V. V., Sergeev A. P.*
UrFU, Ekaterinburg, Russia

THE MAIN METHODS OF INVESTIGATING THE STRUCTURE OF THE TEXT

Abstract

The problem of finding the hidden structures of the test is associated with promising method of attribution. The paper substantiates the choice of methods for the detection and study of such structures in the Russian-language texts. The basic techniques that operate as units of analysis elements that carry meaning, lexical or graphical form, and frequency characteristics of the text. The authors suggest the existence of hidden structures, based on previous results. It is proposed to use frequency analysis as the most adequate to detect hidden structures of the text.

Keywords: text, structure, methodology, semiotics, frequency analysis.

Исследования текста связаны в основном либо с измерением его пространственных характеристик, таких как длина строки, интерлиньяж, размер шрифта, его рисунок, либо с общелингвистическими задачами, когда изучаются смыслонесущие единицы, такие как пред-

ложения, фразы, и т. п. И тот и другой класс изучаемых феноменов не включает в себя скрытые структурные паттерны, не связанные со смыслом текста или с рисунком шрифта.

Структурным исследованиям языка и текста посвящены работы в рамках школы структурализма.

Пражские лингвисты В. Матезиус, Б. Гавранек, Й. Вахек, В. Скаличка устанавливали функциональную дифференциацию традиционного литературного языка. Датские структуралисты Л. Ельмслев, В. Брендаль, Х. Ульдалль (направление глоссематики) хотели создать универсальную и общую для всех теорию в лингвистике.

Американские ученые Ф. Боас, Э. Сепир, Л. Блумфилд в области дескриптивной лингвистики вырабатывали объективные процедуры формального внутреннего анализа того или иного языка при стремлении к научному обоснованию проблем комбинаторики.

Британские структуралисты Дж. Р. Ферс, У. Аллен, М. Керквуд Халлидей, Р. Х. Робинс, У. Хаас, Ф. Р. Палмер работали над решением проблем совместимости идеи создания всеобщей языковой теории, которая была бы применима к конкретному языку, что было бы против принципов глоссематики.

По мнению французских структуралистов, таких как А. Мартине, Г. Гийома и Л. Теньера, язык — это структура, не случайное сочетание слов и звуков, а целое, которое организованно и связанно.

Отечественные исследования в области структуры языка проводились в рамках московско-тартуской школы (Ю. М. Лотман, Б. Ф. Егоров, З. Г. Минц, А. И. Чернов, Б. А. Успенский, В. Н. Топоров, В. В. Иванов, Ю. К. Лекомцев и др.). Представители этой школы объединяются общим исследовательским интересом к проблемам структуры и принципов функционирования систем знаков в сообществе носителей языка, а также общими для всех них структурно-семиотическими методами анализа культурных текстов. Ими также в исследованиях применялись математическое моделирование, элементы теории информации и некоторые статистические методы.

Методологической основой исследований в рамках московско-тартуской школы является тезис де Соссюра о единстве структур всех существующих языков, т. е. сам текст понимается как последовательность знаков, которая может быть проанализирована с позиций именно семиотики, а не семантики.

Из трех составляющих семиотики: семантики, прагматики и синтаксиса, авторов настоящей работы интересует синтаксис как соотношение знаков внутри знаковой системы и, возможно, прагматика, если удается соотнести скрытые структуры и стилистические особенности исследуемых текстов. Семантическая сторона не рассматривается в силу, как уже было сказано, ее конвенциальности, поскольку читатель текста является его интерпретатором, а следовательно, в некоей степени соавтором.

Понятие смысла в разной степени используется во многих распространенных методиках анализа текста. Недостаток подобных методик заключается в отсутствии общепринятого и операционального определения понятия «смысл». Вот некоторые определения этого понятия.

Смысл — значение, которое слово или словосочетание получает в конкретной речевой ситуации (Словарь-справочник лингвистических терминов, Розенталь Д. Э., Теленкова М. А., 1976).

Смысл — идеальное содержание, идея, сущность, значение, предназначение, конечная цель (ценность) чего-либо; целостное содержание какого-либо высказывания, не сводимое к значениям составляющих его частей и элементов, но само определяющее эти значения; разум, разумность. (И. Кондрашин, Глоссарий философских терминов, 2006)

Смысл — содержание изначального самоопределения воли в его отношении к своим инобытийным воплощениям. (Н. Карпицкий, Диалектика человеческого бытия: глоссарий, 1995).

Характеристики пространственной организации текстов и их влияние на чтение и понимание текстовой информации изучаются в рамках исследования удобочитаемости (Weber A. [28], Cohn H. [26], Артемов В. А. [1], Ушакова М. Н. [20], Tinker M. A. [27], Гешев М., Колосов А. [6] и др.). При этом исследуются не столько структура текста как организованного целого (Gestalt), сколько именно удобство чтения, которое в контексте этих исследований соотносимо с понятием юзабилити, которое определяется в системе стандартов ISO как эффективность, результативность, удовлетворенность.

Авторы настоящей работы строят свои исследования на следующих посылках. Во-первых, предполагается, что в тексте существуют скрытые структурные элементы, обнаружение которых возможно специальными методами. Во-вторых, смысл считается конвенцио-

нальным феноменом, поэтому он исключается из рассмотрения из соображений требования объективности исследования.

Цель настоящей работы: анализ литературных источников по вопросам исследования структуры текста для обоснования выбора методики обнаружения скрытых элементов структуры текста.

Структура, текст, структура текста

Безотносительно объекта рассмотрения структура в широком смысле определяется различными авторами следующим образом.

Структура — совокупность устойчивых связей объекта, которые обеспечивают воспроизводимость при меняющихся условиях. В холистическом понимании структура воспринимается как система, которая представляет собой элементы и связи между ними. В другом определении понятия структуры и системы разграничивают. Структура — это внутренняя организация и упорядоченность объекта. В обоих случаях структура подразумевает динамическое и статическое измерения.

Холистическое определение структуры вводит характеристики — целостность и единство. Понятие структуры основывается на диалектическом отношении части и целого. Но целое и части образуют герменевтический круг: части значимы только с точки зрения целого, и наоборот.

Холистическая парадигма доминирует и в структурализме. Леви-Стросс представляет структуру как модель и выделяет четыре холистических критерия:

1) структура обладает свойствами системы. Она состоит из элементов, и изменение одного влечет изменение всех остальных;

2) каждая модель принадлежит к группе преобразований, которые соотносятся с моделью того же семейства, т. е. множество преобразований;

3) обозначенные особенности позволяют предвидеть, как отреагирует модель, если ее элементы будут подвержены модификациям;

4) модель надо сконструировать так, чтобы ее функционирование характеризовало все наблюдаемые факты.

В противоположность холистической парадигме, особенно в математике, распространено представление о структуре как о сетке отношений, связывающей элементы системы. Система в данном случае является совокупностью элементов и сеткой отношений между

ними. Система состоит из совокупности элементов, но они никакого отношения к анализу структуры системы не имеют.

Согласно Ж. Пиаже, структуру можно представить как модель, принятую в лингвистике, математике, логике, физике, биологии и т. д. и отвечающую трем условиям:

1) целостности — подчинение элементов целому и независимость последнего;

2) трансформации — упорядоченный переход одной подструктуры в другую на основе правил порождения;

3) саморегулированию: внутреннее функционирование правил в пределах данной системы. Согласно этому определению, структура тождественна любым системам, в том числе динамическим.

С начала 80-х гг XX в. в социальной теории наблюдается попытка интегративного понимания социальной структуры и социального действия. Тенденция интегративного подхода наиболее четко проявляется в «многомерной социологии» Дж. Александера, в теории «коммуникативного действия» Ю. Хабермаса, в «теории структурации» Э. Гидденса, в когнитивном анализе А. Сикурела и др. Анализ структуры в научных теориях предполагает ответ на вопрос об онтологическом статусе этих структур.

Вопрос об онтологическом статусе структуры в средние века разделял философов на номиналистов и реалистов. Реалисты рассматривали структуру в качестве объективной реальности, существующей независимо от исследователя, а номиналисты не принимали тезис об объективной реальности структуры.

Согласно Леви-Строссу, структуры не составляют часть реальности, а конституируют модели реальности. Структура никакого отношения к реальности не имеет.

Несколько видоизмененный тезис о реальности структуры развивается семиотиками, которые, в соответствии с У. Эко, представляют методологический структурализм, который рассматривает структуру как полезный и необходимый инструмент мышления для целей упрощения различных феноменов с какой-то одной точки зрения.

С точки зрения «теории структурации» Э. Гидденса структуры обладают виртуальным существованием. Структуры характеризуются Гидденсом как вневременные, бессубъектные (цит. по Т. Х. Керимову).

В лингвистически ориентированном структурализме языковые структуры отождествлялись со структурами мышления, и им при-

писывался характер законов, адекватных принципам организации мира, существующего якобы по правилам «универсальной грамматики».

Цв. Тодоров писал: «Эта универсальная грамматика является источником всех универсалий, она дает определение даже самому человеку. Не только все языки, но и все знаковые системы подчиняются одной и той же грамматике. Она универсальна не только потому, что информирует все языки о мире, но и потому, что она совпадает со структурой самого мира» [18].

Определения понятия «текст» столь же разнообразны.

Текст — всякая запечатленная в письменности или в памяти речь, написанные или сказанные кем-нибудь слова, которые можно воспроизвести, повторить в том же виде (толковый словарь Ушакова, 1940).

Текст — написанное высказывание, выходящее за рамки фразы, т. е. являющееся дискурсом и представляющее собой нечто законченное, единое и целое, наделенное внутренней структурой и организацией, соответствующей правилам какого-либо языка. (Философия: энциклопедический словарь; под редакцией А. А. Ивина, 2004).

Текст — законченное, целостное в содержательном и структурном отношении речевое произведение: продукт порождения (производства) речи, отчужденный от субъекта речи (говорящего) и, в свою очередь, являющийся основным объектом ее восприятия и понимания. (Краткий психологический словарь, Л. А. Карпенко, А. В. Петровский, М. Г. Ярошевский, 1998)

Текст реализует структурированно представленную деятельность, а структура деятельности предполагает субъект и объект, сам процесс, цель, средства и результат. Компоненты структуры деятельности отражаются в содержательно-структурных, функциональных, коммуникативных показателях текста.

Принято различать внешнюю (композиционную) и внутреннюю структуры текста. В качестве элементов внешней обычно выступают: предложения, абзацы, разделы и т. д [5].

Под элементами внутренней структуры текста понимают высказывание, фразу, межфразовое единство и т. п.

С семантической, грамматической и композиционной структурой текста связаны его стилевые и стилистические характеристики.

Каждый текст обладает стилистическими особенностями, связан-

ными со структурой текста, которая в свою очередь определяется целью его создания и индивидуальностью автора [16].

Методики анализа текста

1. Интент-анализ

Дает возможность изучать структуру интенции (направленность субъекта на объект) автора по его произведению. Основа метода — экспертное оценивание текста. Это стало причиной того, что полная автоматизация метода не может быть осуществлена [14].

Методики, в которых некоторые этапы этого анализа автоматизированы: Ethnograph, Leximancer, Minnesota Contextual Content Analysis (MCCA).

2. Контент-анализ

В основе метода лежит оценка частотного распределения слов, форм слов и других единиц языка. Итогами являются числовые закономерности и их интерпретация, что позволяет судить о вероятности встречаемости той или иной единицы во всем тексте, частоте встречаемости, относительном и удельном весе.

Методики: Crawdad Desktop, INTEXT, Kwalitan, Protan, Yoshikolder.

3. Фоносемантический анализ

Суть — оценка звучания вне зависимости от содержания. Фонемы и системы их сочетаний подвергаются сопоставлению в определенном слове или даже тексте с их упорядоченными оценками в соответствии со стандартами по ряду биполярных шкал. Анализ позволяет получить вывод о возможности воздействия текста на того, кто его прочитал, а также наглядно увидеть выраженность оценочных шкал семантического пространства в виде профиля. Механизм проработан не до конца, контроль факторов влияния на осмысление текста читателем отсутствует, поэтому многие ученые считают, что его результаты весьма сомнительны и практически не обладают валидностью [14].

Методики: Vaal, DIATON.

4. Дискурс-анализ

Набор методик и приемов интерпретации документов как продуктов речи при определенных общественно-политических и культурно-исторических условиях.

Дает возможность охарактеризовать социальные коммуникации, их различные формальные и второстепенные показатели содержа-

ния. Метод анализа находит широкое применение в исследованиях по социологии и политике, выборочная реализация используется, например, в программе CATPAC.

5. Нарративный анализ

Данный метод основан на сравнении словесных и событийных последовательностей в предложениях. Метод направлен на анализ особых текстов, которые содержат элементы рассказа. Оценивание основано на рассмотрении и оценке конкретных категорий, в частности субъекта, действий. Исследуемые тексты имеют широкий диапазон типов, от класса различных историй и художественных или исторических текстов до обыкновенных статей в периодических изданиях, описывающих те или иные реальные события.

Методики: LIWC, PC-ACE.

6. Экспертная оценка текста

Экспертная оценка текста может производиться относительно информации об авторе текста и для установления временных данных о нем, может быть направлена на определение различных условий написания текста, может служить для выяснения достоверности высказываемых в тексте сведений и установления отдельных признаков.

Комплекс методик, применяемых для проведения вышеупомянутых экспертных оценок:

1) методика изменения текста или его фрагмента без затрагивания его содержания;

2) методика шкалирования с опорой на семантические принципы;

3) методика свободного ассоциативного эксперимента;

4) методика анализа текста в плане его предикативных составляющих.

7. Графематический анализ

Образует фундамент для дальнейшего морфологического и синтаксического анализа с опорой на выделении слов, комплексов цифр, формул.

Служит для:

1) разделения текста на слова или разделители;

2) собирания слов, написанных в разрядку;

3) поиска и выделения фамилий, имен, отчеств, устойчивых оборотов, дат и других числовых данных;

4) поиска и выделения имен, путей и электронных адресов файлов;

5) поиска предложений, абзацев, заголовков разных уровней и примечаний в тексте.

8. Морфологический анализ

Направлен на определение большинства морфологических интерпретаций каждого из слов всего текста, состоящих из лемм, морфологических частиц речи; набора граммем. Реализуется во множестве методик, является основным звеном для других видов текста.

Методики: ATLAS.ti, Textanz, TextArc.

9. Синтаксический анализ

Метод сравнения линейной последовательности лексем языка с его формальной грамматикой. В результате получается синтаксическая модель предложения, которая представлена посредством дерева зависимостей. Финальные итоги синтаксического анализа очень значимы для дальнейших путей переработки текста.

Методики: ProfilerPlus, DictaScope.

10. Семантический анализ

Предназначен для изучения семантической структуры предложения, которая складывается из семантических узлов и семантических отношений. Задачей данного вида анализа является организация этих узлов, образующихся из слов изначального предложения. В фундамент для более четкого определения гипотез в сравнении с составом семантических узлов закладывается информация, которая познается в результате синтаксического анализа. Итоги данного вида анализа можно изобразить в виде семантического графа.

Семантический анализ реализуется посредством различных методик, например, PROTANи в методике T-LAB Tools for Text Analysis, являющаяся компьютерной методикой и производящая три вида анализа: тематический анализ, сравнительный анализ, анализ смежности, а также распознает смысловые паттерны слов и основных идей текста.

11. Частотный анализ

Частотный анализ — это один из способов криптоанализа, который основывается на гипотезе о существовании нетривиального статистического распределения символов и их последовательностей в открытом и шифрованном текстах, которое с точностью до замены символов будет сохраняться в процессе шифрования и дешифрования.

Данный тип анализа базируется на том, что текст включает в себя слова, а слова буквы. Число букв в каждом языке ограничено, и бук-

вы ставятся в некотором определенном порядке. Значимыми показателями текста являются повторяемость букв, пар букв (биграмм) и вообще m-грамм, совместимость букв друг с другом, чередование гласных и согласных и некоторые другие.

Небольшая разница частот в приводимых в разных источниках таблицах раскрывается в том, что частоты очень сильно зависят не только от длины текста, но и еще от его характера. Например, в технических текстах буква Ф очень часто используется в таких словах, как функция, дифференциал, диффузия, коэффициент.

Еще более существенные отклонения от нормы в частоте употребления отдельных букв находятся в таких художественных произведениях, как стихи. Поэтому для более точного определения средней частоты букв необходимо иметь набор разносторонних текстов, которые взяты из всяческих источников. Вместе с тем, как правило, аналогичные изменения не так значимы, и в первом приближении ими можно пренебречь.

Представленный перечень можно расширить за счет методик жанрового, психоаналитического, исторического и других видов анализа. Нетрудно заметить, что ряд методик, таких как интент-анализ, дискурс-анализ, нарративный анализ, экспертная оценка текста, семантический анализ, в качестве единицы анализа используют смыслонесущие элементы или их корреляты (например, интенция). Фоносемантический анализ можно также отнести к этой группе методик в силу его принципиальной конвенциальности, сама идея измерения воздействия текста на читателя предполагает наличие этого самого читателя как активного элемента системы. Другие методики имеют дело с внешними, явными лексическими или графическими формами. К таковым методикам относятся морфологический, графематический, синтаксический и контент-анализ.

Таким образом, для решения задач поиска скрытых структур в тексте наиболее пригодны, по мнению авторов настоящей работы, методики частотного анализа. Во-первых, они не связаны с конвенциальным смыслом текста. Во-вторых, существенные отклонения от нормы в частоте употребления отдельных букв косвенно свидетельствуют о наличии в тексте скрытых структурных элементов.

В рамках метода частотного анализа на сегодняшний день получены результаты, свидетельствующие о наличии в русскоязычных текстах скрытых структурных элементов [21]. Предполагается, что дальнейшие исследования в этом направлении позволят представить

скрытые структуры текста и их элементы в явном виде.

Список литературы

1. Артемов В. А. Технографический анализ суммарных букв нового алфавита // Письменность и революция. M JL. 1933. № 1. С. 58–76.

2. Блауберг И. В., Юдин Э. Г., Становление и сущность системного подхода. М., 1973. гл. IV, § 3.

3. Влавацкая М. В. Комбинаторная лингвистика (постановка проблемы) // Вестник Томского государственного университета. 2011. № 342. С.7–10.

4. Влавацкая М. В. Комбинаторная лингвистика: теоретико-методологические основы зарождения и развития // Научный диалог. 2012. № 3. С. 8–32.

5. Гальперин И. Р. Текст как объект лингвистического исследования. М.: Наука, 1981. 140 с.

6. Гешев М. Я., Колосов А. И. Влияние некоторых технологических факторов набора на удобочитаемость текстов // Научные труды по технологии полиграфического производства. М.: Моск. полигр. ин-т, 1973. № 20. С. 18–21.

7. Гридина Е. А. Анализ алгоритмов автоматического реферирования текста // Восточно-Европейский журнал передовых технологий. 2011. № 2 (51). Т. 3. С.36–38.

8. Деррида Ж. Структура, знак и игра в дискурсе гуманитарных наук // Французская семиотика: от структурализма к постструктурализму / Пер с франц. сост., вступ. ст. Г. К. Косикова. М.: ИГ Прогресс, 2000. С. 407–420.

9. Диковицкий В. В., Шишаев М. Г. Обработка текстов естественного языка в моделях поисковых систем // Труды Кольского научного центра РАН. 2010. № 3. С. 29–34.

10. Иванов В. В. К вопросу о возможности использования лингвистических характеристик сложности текста при исследовании окуломоторной активности при чтении у подростков // Новые исследования. 2013. № 1 (34). С. 42–50.

11. Кузьмина И. В. Феномен языковой избыточности //Вестник Иркутского государственного лингвистического университета. 2011. № 1. С. 142–147.

12. Леонов Н. Н. Математическая социология: структурно-аппроксимационный подход. Мн.: ООО «ФУАинформ», 2002.

13. Лотман Ю. М. Внутри мыслящих миров. Человек — текст — семиосфера — история: избранные статьи: В 3 т. Т. 1. М, 1999. С. 134–136.

14. Митина О. В., Евдокименко А. С. Методы анализа текста: методологические основания и программная реализация // Вестник Южно-Уральского государственного университета. Серия: Психология. 2010. № 40 (216). С. 29–38.

15. Митина О. В., Евдокименко А. С. Формализованные методы исследования текстов: опыт применения к анализу технической документации // Вестник Томского государственного университета. Филология. 2010. № 1. Т. 9. С. 60–69.

16. Пашук Н. С. Психология речи // МИУ — 2010.

17. Родионова О. С. К вопросу о единицах членения текста // Известия высших учебных заведений. Поволжский регион. Гуманитарные науки. 2008. № 1. С. 81–84.

18. Тодоров Цв. Грамматика Декамерона, Mouton, The Hague. Paris, 1969.

19. Уракова Ф. К. Использование языковых единиц разного уровня в построении текста // Известия Российского государственного педагогического университета им. А. И. Герцена. 2008. № 85. С. 258–262.

20. Ушакова М. Н. Новый шрифт для газет // Полиграфическое производство. 1952. № 4. С. 22–23.

21. Филимонов В. В., Живодеров А. А., Горбич Л. Г. Экспрессия и упорядоченность в письменной речи // Известия Уральского федерального университета. Серия 1. Проблемы образования, науки и культуры. 2012. № 3 (104). С. 313–319.

22. Хакимова Е. М. Графические нормы в печатном тексте // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. 2012. № 2 (261). С. 55–59.

23. Щерба Л. В. Языковая система и речевая деятельность: монография. Академия наук СССР. Отд-е литературы и языка. Комиссия по истории филологических наук. Ленинград: Наука. Ленинградское отделение, 1974. 425 с.

24. Яглом А. М., Яглом И. М. Вероятность и информация, М.: Наука, 1973. 513 с.

25. Яцко В. А. Алгоритмы и программы автоматической обработки теста // Вестник Иркутского государственного лингвистического

университета. 2012. № 17 (1). С. 150–161.

26. Cohn. H. Die Hygiene des Auges in den Schulen. Leipzig, 1883.

27. Tinker M. A. Legibility of print. Iowa State University Press, Ames, Iowa, 1963. P. 329.

28. Weber A. Ueber die Augenuntersuchungen in den hoheren schulen zu Darmstadt. Abtheilung fur Gesunheitspflege. Marz, 1881.